On LLM Context Window: Limits and Current Solutions

Gualtiero Marenco Turi

June 28, 2025

1 Introduction

Large Language Models (LLMs), primarily based on the Transformer architecture [1], have demonstrated remarkable abilities in text generation, summarization, question answering, and a wide variety of downstream natural language tasks. A core component of their architecture is the notion of a context window, the fixed size sequence of tokens that the model can attend to at inference time. As of early 2025, the maximum length of this context window has expanded dramatically, from the 2,048 tokens in GPT 3 to over 128,000 tokens in models like GPT 4 turbo, and even 1 million tokens in experimental systems like Gemini 1.5 Pro or Claude 3 Opus.

However, despite these advances in raw sequence length, significant qualitative limitations remain. Models struggle to maintain coherence and factual consistency over long sequences, especially when required to reason across distant parts of the input or retrieve rare, specific information, commonly referred to as the "needle in a haystack" problem. This dichotomy between large nominal capacity and functional effectiveness motivates a closer analysis of both the theoretical and practical constraints of Transformer-based LLMs when working with extended contexts.

At the heart of these limitations lies the fundamental scaling behavior of self-attention, the central mechanism by which LLMs compute dependencies between tokens. The cost of standard attention scales quadratically with input length, limiting the feasibility of arbitrarily extending the context. Furthermore, even when long-context inference is technically supported, models often degrade in performance when confronted with sequence lengths exceeding those seen during pretraining. This behavior is often attributed to positional out of distribution (OOD) generalization failure[2].

The purpose of this work is to investigate the nature of context window limits in LLMs, explain the architectural and training constraints that underlie them, survey the current methods developed to mitigate these limitations and critically assess their effectiveness based on subsequent scientific literature. We also draw attention to tasks for which extended context models still underperform, such as long-form summarization of interdependent content (e.g. literary texts

or dense rulebooks), and explore possible reasons for this.
In doing so, this essay aims to provide an overview of the actual state in an increasingly important research area at the intersection of machine learning scalability, information retrieval, and sequence modeling.

Disclaimer: Throughout this essay, terms such as "reasoning", "inference", "understanding", and "comprehension" are used in line with their widespread usage in the machine learning and NLP literature. However, these terms are anthropomorphic and do not accurately reflect the actual computational behavior of large language models, which operate solely on statistical pattern recognition without any form of semantic grounding or intentionality. Their use here is of covenience and should not be interpreted as endorsing strong cognitive equivalence between LLMs and human thought.

2 The Limits of Contextual Understanding in Large Language Models

While the nominal context windows of Large Language Models (LLMs) have expanded substantially, reaching up to 1 million tokens in experimental configurations, their effective contextual understanding remains fundamentally constrained. These limitations stem from multiple sources: architectural bottlenecks inherent to Transformers, generalization failures beyond training lengths, and practical issues related to how information is encoded, recalled, and integrated across long sequences.

2.1 Architectural Bottlenecks: Quadratic Attention and Positional Encoding

The Transformer architecture, as introduced by Vaswani et al. [1], employs a self-attention mechanism whose computational and memory complexity scales quadratically with input sequence length ($\mathcal{O}(n^2)$). This design inherently limits the feasible size of the context window during both training and inference. Although hardware improvements and memory-efficient variants such as FlashAttention and multi-query attention [3] have alleviated some practical constraints, the core scaling problem persists.
Moreover, the efficacy of attention itself degrades with scale. As the number of tokens increases, attention scores are spread thinly across more entries, resulting in what is often described as *attention dilution*. This leads to a loss in the model's ability to prioritize relevant information, particularly for long-distance dependencies, a critical capability for tasks requiring reasoning across large spans of text.
A related constraint arises from the model's positional encoding scheme. Positional encodings determine how the model understands the order and relative location of tokens. In most LLMs, these are either absolute (e.g. the original

sinusoidal embeddings) or relative (e.g. Rotary Position Embeddings [4]). However, as Jin et al. argue in the SelfExtend work [2], when inference sequences exceed the lengths seen during training, the model encounters positional out of distribution (OOD) inputs, causing degraded and unstable behavior.

2.2 Generalization and Out of Distribution Failures

Empirical studies consistently reveal that LLM suffer from performance drops when tasked with processing sequences significantly longer than those used during training [5]. This issue is particularly evident in language modeling perplexity evaluations and long-context question answering benchmarks.
Such failures are not merely artifacts of computational burden but are deeply tied to inductive biases developed during training. For instance, Pawar et al. [6] provide a comprehensive survey showing that most pretrained LLMs are not naturally extrapolative in the sequence dimension, necessitating explicit architectural or algorithmic interventions for reliable performance in long-context scenarios.

2.3 Limits of Functional Comprehension: The "Infinity vs. Promessi Sposi" Experiment

To illustrate the functional limits of current LLMs, I've consider the difference in summarizing two types of documents: a literary novel such as *I Promessi Sposi* and a dense, rule-based manual such as the *Infinity N5 Rulebook*, a complex tabletop game. While both texts may be similar in length (200+ pages), LLMs were tested and all perform, only metric being human judgement, notably better on the former. This difference arises because novels follow a largely linear narrative structure with recurring entities and thematic coherence—properties well-suited to transformer-based models' inductive biases.
In contrast, rulebooks like that of *Infinity the Game*, beside being more informations-dense, lacks a single narrative spine and instead consist of interdependent logical clauses distributed across the document and references leading to skipping from one page to the other. Understanding such documents requires accurate retrieval of details from earlier sections and dynamic integration with currently processed text—capabilities that degrade with distance due to attention decay and loss of positional fidelity. This limitation reflects what Ratner et al. deems a failure in *parallel context reasoning* [7].

2.4 Trade-Offs and System-Level Limits

It is also important to recognize systemic trade-off in long-context LLM deployment. As Zeng et al. note in their survey on long-context LLM serving [8], increasing context length stresses server memory, bandwidth, and inference latency. These bottlenecks can render long-context support impractical at scale, especially when attempting to serve multiple users concurrently. The CAP principle, balancing consistency, availability, and partition tolerance, emerges as a

framework for reasoning about these trade-offs in LLM environments.
Taken together, these observations suggest that while nominal context length has grown, the effective, reliable span over which LLMs can reason remains bounded by architectural and generalization limits. The next section explores the major approaches proposed to overcome these limits and expand the functional capacity of LLMs in handling long-range contexts.

3 Techniques to Extend Context Window

To mitigate the limitations of standard Transformer-based LLMs in handling long sequences, researchers have developed a range of architectural and algorithmic strategies. These techniques aim to increase the effective receptive field of the model, reduce computational cost, and improve performance on long-context tasks. They can be grouped into five classes: efficient attention mechanisms, memory-augmented models, positional encoding generalization, recurrence or chunk-based processing, and retrieval-augmented generation.

3.1 Efficient Attention Mechanisms

The quadratic complexity of vanilla self-attention motivates the development of *sub-quadratic* attention mechanisms. These include sparse attention, low-rank approximations, and kernelized attention schemes.
Models like Longformer [9] and BigBird [10] implement sparse attention patterns to achieve $\mathcal{O}(n)$ or $\mathcal{O}(n \log n)$ time complexity. Flashattention [3] and its derivatives optimize GPU-level operations to reduce memory overhead and latency during attention computation, but without changing the asymptotic complexity.
More recently, Megadolon [11] introduces a hybrid mechanism using Complex Exponential Moving Averages (CEMA), normalized attention, and autoregressive timestep normalization. This design enables *linear computational and memory complexity* in both training and inference, while outperforming LLAMA2 in downstream tasks despite using a similar parameter budget.

3.2 Memory-Augmented Models

Memory-augmented models extend the context by retaining and reprocessing past activations. Notable examples include Transformer-XL [12], which caches hidden states across segments, and RETRO [13], which retrieves relevant text passages from an external database.
These models break the fixed-context window assumption by treating information outside the current sequence as either learnable recurrence (Transformer-XL) or retrievable context (RETRO). However, they require substantial architectural overhead.

3.3 Generalizing Positional Encoding

The positional encoding scheme of a Transformer governs its ability to generalize across varying sequence lengths. Sinusoidal encodings, as used in the original Transformer, are fixed and deterministic, which limits flexibility. Relative position encodings and more recent methods such as Rotary Position Embeddings (RoPE) [4] offer better extrapolation to longer sequences but still degrade when extrapolated beyond training lengths.

Jin et al. [2] propose *SelfExtend*, a technique that reuses the learned positional encodings by spatial transformation, enabling a form of "stretching" that improves model stability on extended sequences. Experimental results show significant gains on perplexity and QA performance up to 256K tokens, even without architectural modification.

3.4 Chunking, Recurrence, and Windowed Processing

A different strategy involves breaking the long input sequence into smaller, overlapping windows that are processed independently or recurrently. This can be done either during attention (e.g., chunked attention) or at the input level.

MEGA and MEGALODON [11] implement fixed-length chunking during training and inference, combining local EMA-based representations with limited cross-chunk attention to approximate global awareness. This reduces the burden on attention while preserving local contextuality.

Parallel work by Ratner et al. [7] explores *Parallel Context Windows*, where multiple overlapping attention windows are processed in parallel and merged at specific synchronization layers. This enables scalable processing while retaining some degree of global coherence.

3.5 Retrieval-Augmented Generation (RAG)

Another important direction is the use of external retrieval systems to augment model context without increasing the core attention span. In Retrieval-Augmented Generation (RAG) [14], the model issues queries to a document index and integrates retrieved passages into its prompt.

This approach is increasingly used in long-context QA and summarization. However, its performance is gated by the quality of the retriever, its latency, and the model's ability to integrate externally provided information coherently. As Fraga et al.[5] point out, even when relevant information is successfully retrieved, LLMs may still fail to utilize it correctly for multi-hop reasoning or synthesis across disparate retrieved items.

3.6 System-Level Considerations and Serving Techniques

At deployment time, context window extensions are further constrained by infrastructure and serving limitations. Zeng et al. [8] identify the trade-offs involved in long-context serving through the CAP principle: Consistency, Availability, and Partition tolerance. Serving models with context windows above

100K tokens introduces challenges in GPU memory allocation, inference batching, and latency guarantees—especially in shared environments.
To address this, system optimizations (e.g. token-pruning) are actively explored. However most of these solutions benefit speed rather than core comprehension of extended context.

While several promising techniques exist for extending the context capabilities of LLMs, none fully resolve the fundamental bottlenecks without introducing new trade-offs. The next section turns to a detailed analysis of specific phenomena that challenge long-context comprehension, such as attention decay, confabulation, and failure modes in real-world tasks.

4 Pathologies and Challenges in Long-Context Comprehension

Despite the progress made through architectural and algoritmic advances, LLMs still exhibit notable pathologies when operating over long sequences. These issues are often subtle, emerging not from outright failure processing long inputs, but from degradation in reasoning, recall, or consistency. This section explores key failure modes, including attention decay, positional drift, hallucination under long-range compression, and brittle global coherence.

4.1 Attention Decay and Token Interference

As the context length grows, the self-attention mechanism begins to distribute its focus over a larger number of tokens. This leads to what is known as *attention dilution*, where each token receives a smaller fraction of the model's focus. The resulting interference between relevant and irrelevant tokens makes it increasingly difficult for the model to recall precise details from distant parts of the context [7, 5].
In practical terms, this manifests in models recalling incorrect facts, failing to resolve co-references, or responding to questions based on more salient but less relevant information. Attention decay is particularly damaging in tasks where rare but important tokens (e.g. a specific name or date) must be retrieved from distant parts of a long input—a phenomenon known as the *needle in a haystack* problem.

4.2 Positional Drift and Representational Instability

Transformers rely on positional encodings to maintain the order of tokens, but these encodings degrade as they extrapolate beyond the training window. Jin et al. [2] demonstrate that when input lengths exceed training lengths, model behavior becomes increasingly unstable and prone to drift—where the effective representation of positions becomes ambiguous.
This positional drift affects tasks involving structured documents, nested logic,

or long-form generation. For example, in summarizing legal documents or game rulebooks (e.g. the *Infinity N5 Rulebook*), the model often misinterprets hierarchical rules due to confusion over which clause or modifier is in effect—a form of local incoherence that stems from representational degradation.

4.3 Confabulation and Compression Hallucinations

When forced to condense information from very long inputs, LLMs often exhibit compression hallucinations, generating plausible but incorrect details not present in the source. This behavior is especially prevalent in summarization, narrative synthesis, and zero-shot questions answering over extended contexts. As observed in Fraga [5], these hallucinations are not random; they reflect the model's inductive bias toward coherence and frequency over factuality. The model fills gaps in attention or memory using statistically likely continuations, a behavior inherited from its autoregressive training objective.
This failure mode is particularly acute when summarizing documents with high information density but low redundancy. In our earlier comparison, *I Promessi Sposi* being a narrative with recurring motifs—is summarized more accurately than the *Infinity* rulebook, which lacks narrative and contains fine-grained distinctions that do not repeat across sections.

4.4 Inconsistent Long-Range Reasoning

LLMs show significant limitations in reasoning over multiple, widely separated parts of a text. Even when individual pieces of information are correctly recalled, the model may fail to integrate them into a coherent judgment or synthesis.
This failure is well-documented in tasks such as long-context logical reasoning [5, 7]. It is attributed to a combination of weak composition across long sequences and the absence of explicit memory control mechanisms in the standard Transformer.

4.5 Bias Toward Local Context

A recurring empirical finding is that LLMs tend to prefer recent context over distant tokens, even when the latter are more relevant. This bias is a consequence of both the position encoding structure and the training distribution, where shorter contexts dominate.
Pawar et al. [6] highlight this issue in their survey, noting that many long-context models are evaluated on benchmarks that favor localized information retrieval, which can obscure underlying generalization issues. In realistic use cases, reasoning, story comprehension, or long code review, this local bias leads to missed dependencies and superficial understanding.

4.6 LLM Text Understanding and Interpretative fidelity

A fundamental problem in evaluating LLMs on long context tasks, particularly summarization or literary analysis, is that existing metrics rely on surface level alignment with ground truth text rather than any deeper notion of interpretative fidelity. This becomes especially problematic when evaluating outputs over semantically rich and stylistically dense works such as "I Promessi Sposi", where crucial meaning is not explicitly stated, but must be inferred through reader background, intertextual knowledge, or socio-political context.

Current evaluation methods such as ROUGE, BLEU, or even GPT-based judgment models tend to reward verbose regurgitation of salient facts rather than subtle inference or stylistic approximation. This results in a misleading sense of competence, where models appear to "understand" a text based on lexical overlap but miss all relevant subtexts.

This points to a deeper gap: LLMs operate on linguistic form, not semantic depth. While their outputs may correlate with interpretative clarity in some domains, they are demonstrably inadequate in domains where meaning is emergent, contested, or culturally embedded. In this light, evaluating LLM summarization requires a reconsideration of what it means to "understand" a text.

4.7 The Absence of Long Context Benchmarks

A second critical limitation in the current state of LLM research is the lack of widely accepted, robust benchmarks for evaluating performance on long context tasks. Although various synthetic or task specific datasets exist, none offer a comprehensive, real world framework that captures the range of long context challenges.

This absence of reliable benchmarking infrastructure means that claims about long context performance often lack empirical grounding or reproducibility. As Fraga [5] argues, the evaluation landscape is fundamentally inadequate for assessing how models reason, synthesize, or degrade across extended contexts.

Developing rigorous benchmarks is a pressing need for the field. Without them, it remains impossible to compare long context models in a way that is meaningful, standardized, or applicable to real world usage.

These challenges suggest that simply expanding the context window is not sufficient. True long-context comprehension requires structural innovations in memory, attention prioritization, and dynamic reasoning.

5 Conclusions and Future Directions

The ongoing expansion of context windows in large language models has opened new possibilities in long-document understanding, code navigation, dialogue, and intensive tasks. Yet, as this essay has tried to summarize, the increase in sequence length is not linearly matched by gains in comprehension, reasoning,

or factual fidelity. To truly bridge this gap, future LLM architectures must probably move beyond simply scaling.

5.1 Reimagining Memory and Attention Architectures

A growing consensus in the literature suggests that novel memory architectures are essential to break the linear context barrier in a meaningful way. This includes:

- **Persistent State Models:** Techniques like MEGALODON's moving average gated attention [11] or structured state space models [15] offer scalable alternatives to dense attention while preserving long-range signal propagation.
- **Compositional Attention Routing:** Future models may require adaptive attention mechanisms that learn to prioritize specific regions of input based on task demands, in contrast to uniform token attention.

5.2 Training and Evaluation Beyond Static Windows

Current LLM training regimes often employ static-length truncation, with a bias toward short to medium contexts. To support robust generalization training procedures must incorporate perturbation of position encodings and task specific designs [6, 2].
Moreover, evaluation metrics should evolve. Benchmarking LLMs on isolated retrieval or QA tasks fails to capture the complexity of tasks like legal document analysis, long-form reasoning, or multi-document synthesis. As Fraga [5] notes, long-context models must be tested not just for retrieval accuracy but for inferential depth, consistency, and hallucination avoidance.

5.3 Systems Research and Human-AI Interfaces

The challenges of serving long-context LLMs at scale go beyond algorithm design. Zeng et al. [8] emphasize that latency, bandwidth, and memory bottlenecks impose hard constraints on real-time use. Innovations at the systems level will be critical for long-context LLMs to move from the lab to practical deployment.
Equally important are human-facing design principles. When users engage with LLMs over long sessions or documents, the interface must communicate uncertainty and the limits of model recall. An effective long-context LLM is not only computationally capable, but also transparent.

5.4 From Length to Structure

Ultimately, the core insight of this essay is that sequence *length* alone is an inadequate proxy for comprehension. A 500K-token novel and a 200-page rule-

book pose fundamentally different cognitive and architectural challenges. To scale LLMs meaningfully, they must shift focus from context window size to structure.

This shift opens a rich space for interdisciplinary work across machine learning, cognitive science, systems architecture, and human-computer interaction. As we move toward models that serve not just as linguistic mirrors but as structured agents, addressing these structural dimensions seems essential.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[2] Zhiyang Jin, Jiahui Wang, Zhewei Wang, Xinlei Zhang, Qingkai Lin, Peter Liu, Kewei Tu, and Yue Zhao. Selfextend: Scaling sequence length by reusing positional embeddings. *arXiv preprint arXiv:2402.17764*, 2024.

[3] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Re. Flashattention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*, 2022.

[4] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

[5] Natanael Fraga. Challenging llms beyond information retrieval: Reasoning degradation with long context windows. *arXiv preprint arXiv:2403.08911*, 2024.

[6] Saurav Pawar, S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Aman Chadha, and Amitava Das. The what, why, and how of context length extension techniques in large language models – a detailed survey. *arXiv preprint arXiv:2401.08401*, 2024.

[7] Nir Ratner, Yoav Levine, Yonatan Belinkov, Ori Ram, Inbal Magar, Omri Abend, Ehud Karpas, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. Parallel context windows for large language models. *arXiv preprint arXiv:2310.01378*, 2023.

[8] Pai Zeng, Mengwei Xu, Zhenyu Ning, Jieru Zhao, Weihao Cui, and Yizhou Shan. The cap principle for llm serving: A survey of long-context large language model serving. *arXiv preprint arXiv:2403.11228*, 2024.

[9] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[10] Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big bird: Transformers for longer sequences. *Advances in Neural Information Processing Systems*, 33:17283–17297, 2020.

[11] Xuezhe Ma, Xiaomeng Yang, Wenhan Xiong, Beidi Chen, Lili Yu, Hao Zhang, Jonathan May, Luke Zettlemoyer, Omer Levy, and Chunting Zhou. Megalodon: Efficient llm pretraining and inference with unlimited context length. *arXiv preprint arXiv:2403.04792*, 2024.

[12] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[13] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. *arXiv preprint arXiv:2112.04426*, 2022.

[14] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Arturs Kulmizev, Felix Hill, Timo Schick, Wentau Wang, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[15] Albert Gu, Tri Dao Goel, and Alexander Rush. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2206.00870*, 2022.